Oversight Board Comment Submission: AI-Generated Video of Female Campaigner

Digital Rights Foundation
2 July 2026

Since the mass deployment and consumption of AI-enabled tools and applications, policy makers, researchers and human rights activists have been faced with the everchanging challenge of understanding how AI impacts lives across the world and how such harms can be contained without being a threat to legitimate uses for artificial intelligence. In the face of online spaces and digital platforms, this task becomes more pressing than ever before not only because these platforms offer easier access to AI tools, but also because it complicates the risks to individuals and communities alike.

While AI-enabled tools can help people express themselves in new ways and allow them the freedom to present ideas, concepts, and abstract thoughts in more concrete and tangible ways, they can also be exploited for personal gain and put others at danger of being used, misunderstood, harassed, and abused. AI-generated content that mimics real life people, distorts facts and/or attempts at spreading mis/disinformation has dangerous consequences for people, in both online and offline spaces with one slipping into the other. These consequences are especially concerning when they are targeted at individuals from vulnerable positions in society such as women and children.

Multiple research studies have shown that women are disproportionately impacted by AI-assisted harms[1] with creators of AI-generated content exclusively using these tools to invent fake and/or sexualised narratives of women whilst being driven by the intention to enact reputational damage.[2] Similarly, publicly available content altered or edited using AI is also used to to influence public perceptions, evidently to further establish existing societal norms within online spaces as well. Though these digital harms are evidently exacerbated for women living public lives because of their professions (i.e. journalists, politicians or lawyers), reports by the UN show that 58% of young women and girls globally have experienced some form of online harassment online.[3] This report also states that out 99% of deepfakes, 98% are targeted against women while being pornographic in nature. These online attacks against women and young girls highlight a growing threat to women's ability to participate in public life personally, professionally and politically and further jeopardize the status of women in society not only in offline, physical spaces but also in digital ones.

---

[1] https://www.humanrightsresearch.org/post/a-new-form-of-gendered-violence-elon-musk-s-grok
[2] https://journals.sagepub.com/doi/full/10.1177/15248380251384271
[3] https://www.ohchr.org/sites/default/files/documents/issues/women/genderandequality/2025-tool-technology-facilitated-gbv.pdf

In China, for example, parents were commissioning AI-generated videos of women living unmarried, childless lives with fake scenarios seemingly used to fabricate false female narratives to shame and coerce Chinese women by framing their autonomy and choices as social/moral failure.[4] Evidence from India has also shown how AI is strategically used to create sexual imagery of women to suppress their voices, especially when said women engaged in challenging cultural and religious norms in India[5] – a finding made more disturbing when it places minorities, such as Muslim women, at the center who are undoubtedly doubly under threat due to their identity in society.[6] Additionally, in the context of Pakistan, we've seen the use of AI to distort facts in an attempt to discredit and falsely accuse women who participated in Aurat March of having committed anti-state activities during protests putting public facing organizers at risk as well as participating women and gender minorities under threat too.[7]

Women's bodies have long been battlegrounds for communal dominance where men and states stand to gain from controlling stories and perceptions about every-day women. Such abuses are an extension of the role patriarchy plays in societies all over the world - to surveill, punish and ridicule women's bodies and opinions when they assert their own visibility or voice in any setting but especially public ones. This extends to women regardless of what kind of discourse or spaces they interact with. For instance, in January of this year, a photographer from the UK was subjected to online abuse after her personal pictures were altered by AI to be promiscuous. Though a violation on its own, when she chose to speak up about the invasion of privacy and disregard for consent perpetuated by AI-related abuses, users began making even more disturbing sexual images of her.[8]

Given these and many more offences whose incidence is steadily increasing, making AI-assisted, enhanced or enabled tools easily accessible on or through social media platforms such as Facebook, Instagram or X increases the likelihood of such tools being exploited to harm individuals. This was first made apparent in June 2025, when 'nudify' applications were found to be advertised on Meta platforms without any oversight. Though Meta itself had barred companies from putting up such adverts,[9] limiting their reach is only part of the solution to a much bigger problem. The focus should not only be on restricting their reach but also on holding the individuals who exploit these tools to cause harm and spread nefarious content, regardless of how they found out about them or where they downloaded them from.

---

4 https://www.scmp.com/news/people-culture/trending-china/article/3337091/china-parents-buy-ai-clips-regretful-single-women-urge-childless-kids-marry?module=perpetual_scroll_0&pgtype=article

5 https://doi.org/10.1177/00219096241257686

6 https://www.aljazeera.com/features/2026/6/15/looked-so-real-how-ai-is-being-weaponised-against-indias-muslim-women

7 https://rsilpak.org/2024/deepfakes-a-crisis-of-human-rights/

8 https://www.theguardian.com/news/ng-interactive/2026/jan/11/how-grok-nudification-tool-went-viral-x-elon-musk

9 https://about.fb.com/news/2025/06/taking-action-against-nudify-apps/

AI policies need to be woven into the fabric of platform policies by centering the safety of users, opening up effective channels of remedy or recourse, and holding individuals accountable for actions that can and do have real world consequences for people. This means taking into consideration not only when AI-generated content has caused harm to users but also when AI content is generated with the intent to harm, coerce, mislead others, especially when the deception is targeted towards causing someone reputational harm.

Policies that look into what AI-generated content is considered to be harassment, abuse or misleading also need to be adapted to be context-specific. With regions around the world having different cultures, interpretations and societal norms, having a universal, one-size fits all solution puts people at risk of being ignored or unheard in the face of grave challenges and consequences. Broad and anticipatory policies are not the way forward as they run the risk of leaving AI governance pertaining to content moderation and curation largely unregulated.[10]

Having policies that prioritize unique cultural contexts from the Global South as well as the role language plays, can become the make or break behind who gets heard, whose voice is considered important, and who ultimately has access to remedy online. As platforms slowly inch closer to having AI-automated moderation channels whilst simultaneously withdrawing from or backing away from employing independent fact-checking from across the globe, platforms "struggle to effectively moderate content across linguistic diversity of global users."[11]

Another major blind spot within platform policies around AI-generated content is the self-reporting requirement. Currently, Meta's Bullying and Harassment policy requires private adults to self-report unwanted manipulated imagery, while private minors and minor involuntary public figures are protected regardless. Meta's rationale for this requirement is: "...it helps us understand that the person targeted feels bullied or harassed", which is a vague rationale for a highly restrictive caveat to their policy, allowing not just the instant case but many similar cases to slip through the cracks and thus causing irreparable harm to private adults. This requirement to self-report came under criticism[12] in the 2024 "Gender Identity Debate Videos" Oversight Board case, which concerned anti-trans hate content.

In the decision[13], while the Board upheld Meta's decisions to leave up the content, it stated it was "concerned about the self-reporting requirement under the Bullying and Harassment policy and its impact on victims of targeted abuse". The recommendation provided by the Board here

---

10 https://www.cambridge.org/core/journals/german-law-journal/article/digital-media-platforms-and-genai-a-case-for-domainspecific-regulation/84BEBC4EE6C5A8214F11C717C411E00A

11 https://institute.aljazeera.net/en/ajr/article/3419

12 https://www.techpolicy.press/will-the-oversight-board-finally-get-meta-to-enforce-its-own-policies-on-anti-trans-hate-and-disinformation/

13 https://www.oversightboard.com/decision/bun-1ynnk264/

was to "Allow users to designate connected accounts, which are able to flag potential Bullying and Harassment violations requiring self-reporting, on their behalf". This recommendation should be further extended by removing the requirement to self-report, as the rationale is too frail when balanced against the very real harms that keeping harmful content up causes. In many cases, the private individual might not even be aware of the content for a long period of time, or until they suffer direct reputational harm from it, or might be so strongly mentally affected by the content that they are simply unable to self-report.

Without proper rules and regulations in place for penalizing wrongful and harmful content, the internet creates barriers to accurate, efficient and sometimes life changing information sharing mechanisms; barriers which are more often than not mere reflections of how society already operates for vulnerable, marginalized people. Women's access to accurate and detailed health information has been a topic of discussion in research and academia circles for a long time. Historically, women have been unable to find medical information and advice on female health, especially when it comes to menstruation, reproduction or other gender-specific health conditions.

In addition to the lack of importance given to female bodies and health, this vacuum of information has been a direct consequence of taboos and social norms in patriarchal societies across the world where talking about the female body and its processes has been considered vulgar, coarse and obscene. These problems are further exacerbated by social media where content that attempts to address these gaps of information has been flagged, suspended or censored online for being against platform content policies.[14]

According to a research conducted by the Digital Freedom Fund, social media platforms are systematically allowing posts and content discussing men's health to stay up while content relating to women's health is misclassified as "sexual or adult" and is subsequently reported, suspended, and removed.[15] When such content also refers to diagrams or pictures of the female body are flagged as pornography without taking into account the context under which these images are being shared. Another barrier to access of vital and life-changing information is the prevalence of mis/disinformation when it comes to health services that women can approach. Research conducted by MSI in the Global South shows how medically inaccurate information and anti-choice rhetoric are more likely to be promoted on social media platforms.[16] In other instances, fake clinic information is also circulated online putting women and girls at the risk of serious life-threatening danger.

---

[14] https://www.context.news/big-tech/advocates-say-online-erasure-of-womens-health-is-dangerous

[15] https://digitalfreedomfund.org/case-studies/gender-based-censorship-of-sexual-and-reproductive-health-content-by-online-platforms/

[16] https://www.msichoices.org/latest/digital-disparities-the-global-battle-for-reproductive-rights-on-social-media/

Further evidence from this Oversight Board case itself, individuals who choose to talk about such important issues and raise awareness on women reproductive health are subjected to blatant harassment and are digitally surveilled for their work.[17] Oftentimes, this can lead to dangerous circumstances where their safety comes under risk and can also hinder their ability to conduct work and support women.

That being said, the newly selected Oversight Board case is not just about one misleading video. It raises a deeper product-governance problem around when a woman is turned into a viral object of ridicule through AI manipulation, the harm is produced not only by the falsity of the media, but by the way platform design accelerates visibility, participation and repeat exposure.

The closest recent example of this is the Board's 2026 decision in a related case involving an AI-generated sexualized impersonation video on Instagram. In the decision, the Board held that AI-generated sexualized impersonation of a real person should be treated as non-consensual by default, and rejected age-gating as an adequate response. The board recommended connected-account reporting, a dedicated reporting category and broader signals of lack of consent. The earlier decision is highly relevant because the campaigner case exposes a similar gap in Meta's logic. The company treats manipulated imagery as actionable when it alters how a person looks, but not when it fabricates what she is doing, saying, endorsing or representing. The policy misses a major share of real-world abuse and in practice, the reputational and participatory harm can come from false attribution of speech, conduct, affiliation, expertise, sexuality or morality just as much as from explicit nudification. This is especially true when the target is a non-public figure or a woman in civic life, because the objective is often humiliation, silencing and the triggering of mass commentary rather than deception alone.[18]

Social media companies reduce mass harassment most effectively when they treat it as a systems-design problem rather than a queue of individual abusive posts. The first design lever is recommendation and ranking. A large-scale randomized experiment on an Indian TikTok-like platform found that replacing algorithmic ranking with random content delivery reduced exposure to toxic posts by 27 percent.[19] Separate audit research on TikTok found that interest-aligned content is strongly amplified and often rapidly reinforced within roughly the first 200 videos watched, while content diversity declines as the system learns user preferences.[20]

---

[17] https://womenslinkworldwide.org/en/defending-reproductive-justice-in-digital-spaces

[18] https://www.oversightboard.com/decision/ig-4ueqal8u/?utm_campaign=47261761-2025x%20-%20AI%20Sexualised%20Video%20Case&utm_medium=email&_hsenc=p2ANqtz--onpxxiRNzej2z9cdZxr88YOK7C3ArJVQkbZPZMz4AUI1yCZJPLKyvKQ8poU15RG2lluGKzFZzyri1wwld236mR6U_ahEf1LEppiJ488pnudnbtQM&_hsmi=425118118&utm_content=425118118&utm_source=hs_email

[19] https://www.researchgate.net/publication/389714056_Hate_in_the_Time_of_Algorithms_Evidence_on_Online_Behavior_from_a_Large-Scale_Experiment

[20] https://www.researchgate.net/publication/400682000_Dynamics_of_algorithmic_content_amplification_on_TikTok

For harassment, the implication is therefore straightforward that once ridicule, body shaming or misogynistic commentary starts performing well, engagement-driven ranking can turn a localized attack into a platform-wide practice.

Moreover, posts become pile-ons because platforms make copying, reposting, quote-posting, forwarding, remixing and replying extremely easy. Research on WhatsApp's forwarding limits found that virality restrictions can significantly delay the spread of false content, even if such measures are not sufficient on their own for highly viral campaigns.[21] The same principle applies to cases of harassment where slowing quote-posts, limiting repeated forwards or pausing broad remix functions buys time for review and interrupts mob formation. In other words, platform design should make the first abusive post easier to contain, not easier to scale.

Harassment is also often not visible if moderators only review one post at a time. Platforms should look for surge patterns where sudden influxes of non-followers, high-volume replies in a short window, repeated slurs or appearance-based remarks, sudden repost cascades or multiple uploads targeting the same person. Once such a surge is detected, the platform should automatically shift into a higher-safety mode where a cap on the number of replies per hour, turn off quote-posting, suppress recommendation surfaces and route the target's case to expedited human review. Ofcom's guidance for protecting women and girls has already pointed toward response limits, prompts that interrupt harmful posting, time-outs for repeat abusers and bulk block or mute tools as practical measures.[22] Bluesky's quote-detach feature is a smaller but useful example of a design intervention aimed specifically at dogpiling.[23]

Also, if abuse can earn reach, followers or money, the platform is subsidizing harassment. Ofcom's guidance also recommends preventing misogynistic users from monetizing posts, which matters because many pile-ons are opportunistic rather than ideological and are driven by attention markets.[24] Platforms should therefore remove content from monetization that attacks individuals, discount it in ranking and suspend repeated offenders from recommendation surfaces even before a final account penalty is imposed.

Again in the case of harassment and abuse online victims should not have to block hundreds of accounts one by one while a pile-on is unfolding. Platforms should offer one-tap defensive tools with bulk blocking or muting, temporary locking of replies, priority notification filters and the ability to hide or collapse hostile replies automatically. These are not just comfort features. Survey research in the UK shows women are significantly less comfortable expressing political

---

[21] https://www.researchgate.net/publication/381058794_Don't_Break_the_Chain_Measuring_Message_Forwarding_on_WhatsApp

[22] https://www.mishcon.com/news/online-safety-ofcom-publishes-guidance-to-tech-firms-to-tackle-online-harms-against-women-and-girls

[23] https://www.theverge.com/2024/8/29/24231414/bluesky-anti-toxicity-features-detach-posts-from-quotes

[24] https://www.taylorwessing.com/en/insights-and-events/insights/2026/01/rd-ofcoms-guidance-on-online-safety-for-women-and-girls-and-uk-government-strategy

opinions online than men[25], and longitudinal research on toxic replies on X/Twitter shows that abuse changes user behavior through avoidance, countermeasures, and other defensive responses.[26] Platform design, therefore, directly shapes whether women remain able to speak in public or are pushed into silence.

Platforms should adopt a broader definition of unwanted manipulated imagery that covers not only alterations to a person's appearance, but also AI-generated manipulation of their voice, speech, actions, identity, context or apparent affiliations. Given the growing prevalence of AI-generated impersonation, platforms should presume a lack of consent where manipulated content is used to humiliate, sexualize or misrepresent individuals, particularly women and marginalized groups, rather than placing the burden on victims to prove harm.

Platforms should also move beyond self-reporting as the primary trigger for enforcement. Survivors may be unaware of abusive content, having left the platform, or be unable to report due to trauma. Instead, trusted third parties, including designated contacts, civil society organizations and legal representatives, should be able to report on behalf of victims. Dedicated reporting channels for AI-generated manipulated media should be introduced, supported by rapid human review, temporary de-amplification while cases are assessed, and hash-matching tools to prevent re-uploads. While AI labels and provenance metadata may assist moderation, they should complement, not replace, effective removal, de-amplification and survivor-centered enforcement measures.

[25] https://iknowpolitics.org/en/learn/knowledge-resources/three-four-women-not-comfortable-expressing-political-opinions-online
[26] https://arxiv.org/abs/2210.13420